Non-parametric shrinkage mean estimation for arbitrary quadratic loss functions and unknown covariance matrices

Cheng Wang [*1,3], Tiejun Tong[2], Longbing Cao[3], and Baiqi Miao[1]

[1]Department of Statistics and Finance, University of Science and Technology of China, Hefei, Anhui 230026, China
[2]Department of Mathematics, Hong Kong Baptist University, Hong Kong
[3]Advanced Analytics Institute, University of Technology, Sydney, NSW 2007, Australia

December 27, 2012

Abstract

In this paper, a shrinkage estimator for the population mean is proposed under arbitrary quadratic loss functions with unknown covariance matrices. The new estimator is non-parametric in the sense that it does not assume a specific parametric distribution for the data and it does not require the prior information on the population covariance matrix. Analytical results on the improvement of the proposed shrinkage estimator are provided and some corresponding asymptotic properties are also derived. Finally, we demonstrate the practical improvement of the proposed method over existing methods through extensive simulation studies and real data analysis.

Keywords: High-dimensional data; Shrinkage estimator; Large p small n; U-statistic.

[*]cescwang@gmail.com

1   Introduction

High-throughput molecular technologies that enable researchers to collect and monitor information at the genome level have revolutionized the field of biology in the past fifteen years. These data offer an unprecedented amount and diverse types of data that reveal different aspects of the biological processes. One such example is microarray data, where the expression levels of thousands of genes are measured simultaneously from each sample. These data have motivated the development of reliable biomarkers for disease subtypes classification and diagnosis, and for the identification of novel targets for drug treatment. Due to the cost and other experimental difficulties such as the availabilities of biological materials, it is common that high-throughput data are collected only in a limited number of samples. They are often referred to as high-dimension, low-sample-size data, or "large p small n" data where p is the number of genes or dimensions and n is the sample size. High-dimensional data pose many challenges to traditional statistical and computational methods. Specifically, due to the small n, there are more uncertainties associated with standard estimations of parameters such as the mean and variance estimations. As a consequence, statistical analyses based on such parameter estimation are usually unreliable.

To obtain more accurate parameter estimation some statistical methods, such as shrinkage, may yield better results. In the last decade, the researchers have proposed quite a few shrinkage-based methods to enrich the literature in shrinkage methodology under the "large p small n" setting, with particular interests on the variance estimation (Tusher *et al.*, 2001; Baldi and Long, 2001; Smyth, 2004; Cui *et al.*, 2005; Tong and Wang, 2007; Tong *et al.*, 2012b) and on the covariance matrix estimation (Ledoit and Wolf, 2004a; Schäfer and Strimmer, 2005; Pourahmadi, 2011; Cai and Yuan, 2012). Apart from the progress made on the variance and covariance matrix estimations, some attention has been paid recently to the estimation of the population mean μ under the "large p small n" setting (Hwang and Liu, 2010; Tong *et al.*, 2012a). An accurate estimate of μ is desired in many areas of statistical analysis, e.g., in linear discriminant analysis (Anderson, 2003), diagonal linear discriminant analysis (Dudoit *et al.*, 2002), Markowitz mean-variance analysis (Markowitz, 1952; El Karoui, 2010) and so on.

Under the assumption that μ is sparse, Shao *et al.* (2011) proposed a consistent estimator for μ under some regular conditions. However, in many real problems, there is often little prior information on μ and it may not nec-

essarily have a sparse structure. In such situations, the shrinkage estimation of μ can be applied. Shrinkage estimation starts with the amazing result of James and Stein (1961) that the commonly used sample mean of a normal distribution is inadmissible and can be improved by shrinkage estimators. We refer to them as James-Stein type estimators. Since then, there is a large body of literature in shrinkage estimation including Baranchik (1970), Efron and Morris (1973), Lin and Tsai (1973), Berger *et al.* (1977), Gleser (1986), Fourdrinier *et al.* (2003), and etc. In the literature, most existing methods either assumed that the covariance matrix Σ_p is known or assumed that there exists an estimator of Σ_p that is invertible. As a common practice, if the sample covariance matrix S_n is used to estimate Σ_p, the sample size is required to be larger than the dimension, i.e. $n > p$, to avoid the singularity. Note that, however, for high-dimensional data it is common that p is much larger than n. Therefore, the traditional shrinkage methods can not be applied to analyze high-dimensional data directly.

To overcome the singularity problem, Tong *et al.* (2012a) proposed a new shrinkage estimator for μ by assuming that Σ_p has a diagonal structure. This assumption is equivalent to assuming that the genes are independent of each other. Though it may not be realistic, we note that the diagonal assumption on Σ_p has been made frequently in different aspects of high-dimensional data analysis, e.g., in high-dimensional classification (Fan and Fan, 2008; Pang *et al.*, 2009), in shrinkage estimation of variances (Tong and Wang, 2007; Hwang *et al.*, 2009) and the references therein. Needless to say, the diagonal assumption is very restrictive. Recently, Ramey (2012) and Fan *et al.* (2012) pointed out that the diagonal related discriminant classifiers in Dudoit *et al.* (2002), Pang *et al.* (2009) and Tong *et al.* (2012a) can be suboptimal in real data sets classification owing to the information loss in off-diagonal elements. In addition, the shrinkage method in Tong *et al.* (2012a) requires the data to be Gaussian distributed through a Bayesian model. These restrictions have largely limited the usage of existing shrinkage methods in high-dimensional data. It is also worth pointing out that the research so far has concentrated on the modelling and little is known about the theoretical properties of various shrinkage estimators.

Inspired by Ledoit and Wolf (2004b), in this paper we consider the shrinkage estimation for μ under arbitrary quadratic loss functions with unknown non-diagonal covariance matrix. The new estimator is non-parametric in the sense that it does not assume a specific parametric distribution for the data and it does not require the prior information on covariance matrix Σ_p. We

will demonstrate by both theoretical and empirical studies that the proposed estimator has good properties for a wide range of settings. We will also show that the proposed method is better than the sample mean and the existing shrinkage methods even under a diagonal covariance matrix assumption. The rest of the paper is organized as follows. Section 2 introduces the theoretical optimal shrinkage estimation under quadratic risks. Section 3 develops a data-driven shrinkage estimator and derives the asymptotic properties of the proposed estimator. We then conduct simulation studies using simulated data in Section 4 and using real data in Section 5 to evaluate the proposed optimal shrinkage estimator and compare it with existing shrinkage methods. Finally, we conclude the paper in Section 6 and provide the technical results in the Appendix.

2   Methodology

Let $X_1, \cdots, X_n$ be independent and identically distributed (i.i.d.) observations satisfying the multivariate model

$$X_i = \Sigma_p^{1/2} \epsilon_i + \mu, \qquad i = 1, \cdots, n, \tag{2.1}$$

where μ is a p-dimensional vector, Σ_p is a positive definite matrix and the random errors in $(\epsilon_{ij})_{p \times n} = (\epsilon_1, \cdots, \epsilon_n)$ are i.i.d. with zero mean, unit variance and finite fourth moment. Note that model (2.1) is assumed the same as those in Bai and Saranadasa (1996) and Chen *et al.* (2010). In this paper, we do not assume that the data follow a multivariate normal distribution with mean μ and covariance matrix Σ_p. Given model (2.1), we consider to estimate μ under the following quadratic loss function (Berger, 1976; Berger *et al.*, 1977; Gleser, 1986),

$$L_Q(\delta) = n(\delta - \mu)'Q(\delta - \mu)/\text{tr}(Q\Sigma_p), \tag{2.2}$$

where $\delta = \delta(X_1, \cdots, X_n)$ is the estimator of μ, Q is a known positive definite matrix, and $\text{tr}(A)$ stands for the trace of matrix A. Note that for the standard sample mean $\bar{X} = (1/n) \sum_{k=1}^{n} X_k$, the risk function is $E[L_Q(\bar{X})] = 1$.

In the special case when $X_1, \cdots, X_n$ are multivariate normal distributed, James and Stein (1961) showed that

$$\delta_{JS} = (1 - \frac{p-2}{n\bar{X}'\bar{X}})\bar{X} \tag{2.3}$$

dominates $\bar{X}$ for any $p > 2$ under the assumption that $\Sigma_p = Q = I_p$. This result was then extended by Baranchik (1970) to $\Sigma_p = \sigma^2 I_p$ with σ^2 unknown, and by Efron and Morris (1973) to a Bayesian estimator. For a general unknown Σ_p, the James-Stein estimator has the form (Lin and Tsai, 1973; Berger, 1976; Berger $et\ al.$, 1977; Gleser, 1986; Fourdrinier $et\ al.$, 2003)

$$\delta_{JS} = (I - \frac{r(Q, S_n^{-1}, \bar{X})}{\bar{X}'S_n^{-1}\bar{X}})\bar{X}, \tag{2.4}$$

where S_n is the sample covariance matrix which is defined as

$$S_n = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X})(X_k - \bar{X})'.$$

To guarantee S_n is invertible, $n > p$ is necessary which means the method is not applicable for "large p, small n" data.

To overcome the singularity problem, Tong $et\ al.$ (2012a) considered a special situation where Σ_p is diagonal. Specifically, under the loss function with $Q = \Sigma_p^{-1}$ they constructed a hierarchical Bayesian model and then proposed the following shrinkage estimator,

$$\delta_T = (1 - \frac{(p-2)(n-1)}{n(n-3)\bar{X}'D_n^{-1}\bar{X}})\bar{X} \tag{2.5}$$

where $D_n = \text{diag}(S_n)$ is the diagonal sample covariance matrix. Other related works for a diagonal Σ_p and a diagonal Q assumptions include Berger and Bock (1976) and Shinozaki (1980). Whereas for an arbitrary Q with non-diagonal Σ_p, it remains a challenging yet unanswered question under the "large p small n" setting. To address this question, we consider to estimate μ by a linear combination of $\bar{X}$ and $e = (1, \cdots, 1)'$,

$$\delta = \alpha\bar{X} + \beta e.$$

The following theorem derives the optimal shrinkage coefficients for model (2.1) under the quadratic loss (2.2) with an arbitrary Q.

Theorem 2.1 *Consider the optimization problem,*

$$\min_{\alpha, \beta} E(\delta - \mu)'Q(\delta - \mu) \quad \text{s.t.} \quad \delta = \alpha\bar{X} + \beta e, \tag{2.6}$$

5

where the coefficients α and β are non-random. The optimal shrinkage estimator is given as $\mu^* = \alpha^* \bar{X} + \beta^* e$ where

$$\alpha^* = \frac{\mu'Q\mu - \frac{(e'Q\mu)^2}{e'Qe}}{\mu'Q\mu + \frac{1}{n}\text{tr}(Q\Sigma_p) - \frac{(e'Q\mu)^2}{e'Qe}},$$

$$\beta^* = \frac{\frac{1}{n}\text{tr}(Q\Sigma_p)}{\mu'Q\mu + \frac{1}{n}\text{tr}(Q\Sigma_p) - \frac{(e'Q\mu)^2}{e'Qe}} \frac{e'Q\mu}{e'Qe},$$

and the corresponding risk of μ^* is

$$E(L_Q(\mu^*)) = \frac{(\mu - \frac{e'Q\mu}{e'Qe}e)'Q(\mu - \frac{e'Q\mu}{e'Qe}e)}{(\mu - \frac{e'Q\mu}{e'Qe}e)'Q(\mu - \frac{e'Q\mu}{e'Qe}e) + \frac{1}{n}\text{tr}(Q\Sigma_p)}. \tag{2.7}$$

Note that the proposed shrinkage estimator can accommodate any shift of the grand mean, including the shift from μ to $\mu + ce$ where c is a constant. This is a similar idea as that in Lindley (1962) where the author shrunk the observations to grand mean rather than to the origin. Also in Tong *et al.* (2012a), the authors applied their shrinkage method to the grand mean and so the final estimator was a linear combination of two different components. By Theorem 2.1, however, we point out that the method in Lindley (1962) is not applicable for arbitrary Q. For this point, we will expain in the simulation study an example where the grand mean is zero but $e'Q\mu \neq 0$.

3   Data-driven shrinkage estimators for population means

Note that the shrinkage coefficients α^* and β^* are unknown and need to be estimated in practice. In this section, we propose to estimate them by U-statistics, motivated from Chen *et al.* (2010), Cai and Ma (2012) and Li and Chen (2012). Specifically, we estimate α^* and β^* by

$$\hat{\alpha}^* = \frac{Y_{1,n} - Y_{3,n}}{Y_{1,n} + Y_{2,n} - Y_{3,n}} \quad \text{and} \quad \hat{\beta}^* = \frac{Y_{2,n}}{Y_{1,n} + Y_{2,n} - Y_{3,n}} Y_{4,n}$$

where

$$Y_{1,n} = \frac{1}{p(n-1)} \sum_{i \neq j} X_i' Q X_j,$$

$$Y_{2,n} = \frac{1}{np} \left(\sum_{k=1}^{n} X_k' Q X_k - \frac{1}{n-1} \sum_{i \neq j} X_i' Q X_j \right),$$

$$Y_{3,n} = \frac{1}{p(n-1)e'Qe} \sum_{i \neq j} e' Q X_i X_j' Q e,$$

$$Y_{4,n} = \frac{1}{ne'Qe} \sum_{k=1}^{n} e' Q X_k.$$

The resulting estimator of μ is then $\hat{\mu}^* = \hat{\alpha}^* \bar{X} + \hat{\beta}^* e$. To derive the asymptotic properties of the proposed estimator, we need the following regularity condition.

Assumption 3.1 *There is a constant c_0 (not depending on p or n) such that*

$$c_0^{-1} \leq \text{all eigenvalues of } \Sigma_p \text{ and } Q \leq c_0$$

Under Assumption 3.1, we have $\text{tr}(\Sigma_p Q)/p = O(1)$. In this work, $o(1)$ denotes a sequence of random variables that converges to zero and $O(1)$ is short for a sequence that is bounded. Similarly, $o_p(1)$ and $O_p(1)$ are notations in probability. For more details, one may refer to Van der Vaart (2000). Let

$$\pi_1 = E(\bar{X} - \mu)' Q (\bar{X} - \mu) = \frac{1}{n} \text{tr}(Q \Sigma_p),$$

$$\pi_2 = (\mu - \frac{e'Q\mu}{e'Qe} e)' Q (\mu - \frac{e'Q\mu}{e'Qe} e).$$

The following theorems establish the rates of convergence for the proposed estimators and for the loss function.

Theorem 3.1 *Under Assumption 3.1,*

$$Y_{1,n} = \frac{n}{p} \mu' Q \mu + O_p(\frac{1}{\sqrt{p}}), \qquad Y_{2,n} = \frac{1}{p} \text{tr}(\Sigma_p Q) + O_p(\frac{1}{\sqrt{np}}),$$

$$Y_{3,n} = \frac{n(\mu' Q e)^2}{pe'Qe} + O_p(\frac{1}{p}), \qquad Y_{4,n} = \frac{e'Qu}{e'Qe} + O_p(\frac{1}{\sqrt{np}}).$$

7

Further, we have

$$\hat{\alpha}^* = \alpha^* + O_p(\frac{1}{\sqrt{p}}) \quad \text{and} \quad \hat{\beta}^* = \beta^* + O_p(\frac{1}{\sqrt{np}}) + O_p(\frac{e'Qu}{p^{3/2}(\frac{n}{p}\pi_1 + \frac{n}{p}\pi_2)}).$$

Therefore, $\hat{\alpha}^* - \alpha^* \xrightarrow{p} 0$ and $\hat{\beta}^* - \beta^* \xrightarrow{p} 0$ as $p \to \infty$ and $p^{-3/2}e'Qu \to 0$, where $\xrightarrow{p}$ denotes convergence in probability.

Theorem 3.2 *Under Assumption 3.1, under the "large p small n" setting the loss function of the shrinkage estimator $\hat{\mu}^*$ is*

$$L_Q(\hat{\mu}^*) = \frac{\pi_2}{\pi_1 + \pi_2} + O_p(\frac{1}{\sqrt{p}}). \tag{3.8}$$

By Theorem 3.2, we note that $\hat{\mu}^*$ behaves at least as well as $\bar{X}$ when p is large. The explicit improvement of $\hat{\mu}^*$ over $\bar{X}$ depends on π_1 and π_2. As in Ledoit and Wolf (2004b), we define the percentage relative improvement in average loss (PRIAL) over the sample mean as

$$\text{PRIAL} = \frac{\text{E}(\bar{X} - \mu)'Q(\bar{X} - \mu) - (\hat{\mu}^* - \mu)'Q(\hat{\mu}^* - \mu)}{\text{E}(\bar{X} - \mu)'Q(\bar{X} - \mu)}. \tag{3.9}$$

We then have the following corollary.

Corollary 3.1 *Let $s_n = \frac{n}{p}(\mu - \frac{e'Q\mu}{e'Qe}e)'Q(\mu - \frac{e'Q\mu}{e'Qe}e)$. As $p \to \infty$ we have*

(I) If $s_n \to 0$, PRIAL $\xrightarrow{p} 1$;

(II) If $s_n \to C_0$, PRIAL $\xrightarrow{p} C_1 \in (0, 1)$;

(III) If $s_n \to \infty$, PRIAL $\xrightarrow{p} 0$.

Therefore, the shrinkage estimator $\hat{\mu}^*$ always performs better than $\bar{X}$ under the loss function (2.2) when s_n is finite. In the extreme case when $s_n \to \infty$, $\hat{\mu}^*$ behaves similarly as $\bar{X}$.

4 Simulation studies

In this section, we conduct simulation studies to evaluate the performance of the proposed shrinkage estimator $\hat{\mu}^*$ and compare it with the following four estimators: the sample mean $\bar{X}$, the James-Stein estimator δ_{JS} in Baranchik (1970), the Berger-Bock estimator δ_B in Berger and Bock (1976), and the Tong et.al. estimator δ_T in Tong et $al.$ (2012a).

Note that the existing competitors for comparison, δ_{JS}, δ_B and δ_T, only work on a diagonal covariance matrix under the "large p small n" setting. Whereas for the proposed estimator $\hat{\mu}^*$, it works for both diagonal and non-diagonal covariance matrices. Thus for a meaningful comparison, we will consider the quadratic loss function in Tong et $al.$ (2012a). Specifically, by letting Q be diagonal and let $Q^{-1} = \text{diag}(\Sigma_{\text{p}})$, we have the following loss function,

$$L(\delta) = \frac{n}{p}(\delta - \mu)'[\text{diag}(\Sigma_{\text{p}})]^{-1}(\delta - \mu), \tag{4.10}$$

where the constant n/p is applied to guarantee that $E[L(\bar{X})] = 1$. In applications, Q will be estimated from the diagonal elements of the sample covariance matrix.

We simulate $X_1, \cdots, X_n$ independently from a p-dimensional multivariate normal distribution with mean μ and covariance matrix Σ_p. For μ, we consider two options:

(a) Let $\mu_1 = (\mu_{11}, \cdots, \mu_{1p})'$ where $\mu_{11}, \cdots, \mu_{1p}$ are i.i.d from $N(0, \tau^2)$; and

(b) Let $\mu_2 = (\mu_{21}, \cdots, \mu_{2p})'$ where $\mu_{2k} = \tau$ for $k \leq p/2$ and $\mu_{2k} = -\tau$ for $k > p/2$.

In both options, we consider $\tau = 0.5$ and 1 to represent different levels of mean heterogeneity. For Σ_p, we consider three covariance matrices:

(1) Σ_1 is diagonal with 20% of population eigenvalues being equal to 1, 40% begin equal to 3 and 40% being equal to 10;

(2) $\Sigma_2 = \Sigma_1^{1/2}\Sigma_0\Sigma_1^{1/2}$ where $\Sigma_0 = (\sigma_{ij})_{p \times p}$ and $\sigma_{ij} = \rho^{|i-j|}$ for $1 \leq i, j \leq p$;

(3) $\Sigma_3 = \Sigma_1^{1/2}\Sigma_{00}\Sigma_1^{1/2}$ where $\Sigma_{00} = (\sigma_{ij})_{p \times p}$ and $\sigma_{ij} = 1$ for $i = j$, $\sigma_{ij} = \rho$ for $i \neq j$.

9

We let ρ range from 0.1 to 0.9 for Σ_2 and from 0.1 to 0.5 for Σ_3 to represent four different levels of dependent structure.

The first simulation study is to evaluate the performance of $\hat{\mu}^*$ with existing methods when $\Sigma_p = \Sigma_1$, i.e., when the covariance matrix is diagonal. Let $p = 100$ throughout the simulations. We consider $n = 10, 25, 50$ and 100, to represent different levels of sample sizes. Table 4 reports the average risks of the estimators under various settings, based on 10,000 simulations. First of all, we observe that all shrinkage methods have a smaller risk than the sample mean $\bar{X}$. This shows that for high-dimensional data, the shrinkage estimators do improve the standard estimation. Among the shrinkage estimators, δ_T and $\hat{\mu}^*$ are among the best in most settings. The James-Stein estimator δ_{JS} is not very compatible because it is restricted to a common variance assumption, and δ_B is only applicable for large sample sizes. Finally, for δ_T and $\hat{\mu}^*$, we note that they perform similarly when $\mu = \mu_1$, and $\hat{\mu}^*$ is better by a large margin than δ_T when $\mu = \mu_2$. In addition, when the mean heterogeneity increases from $\tau = 0.5$ to $\tau = 1$, the improvement of $\hat{\mu}^*$ over $\bar{X}$ decreases which is consistent with Corollary 3.1. We also observe that the improvements of the shrinkage estimators over the sample mean become smaller when n becomes larger. This is meaningful since for the large sample size scenario, the mean estimation itself is good enough and it is no longer necessary to borrow information from others to improve the estimation.

The second simulation study is to evaluate the performance of $\hat{\mu}^*$ with existing methods when the covariance matrix is non-diagonal. This is to investigate the impact of the correlation coefficient ρ on the performance of the estimators. To achieve this, we plot in Figure 1 the average risks of the estimators for covariance matrices Σ_2 and Σ_3 respectively, based on $100,000$ replications. To save space, we only present the results for $p = 100$, $n = 20$, $\mu = \mu_1$ and $\tau = 0.5$; whereas the comparison patterns for other combination settings remain the similar. From the plots, it is evident that the proposed $\hat{\mu}^*$ provides a smaller average risk than the other estimators in most settings, no matter if ρ is small or not. We also note that (i) all the shrinkage estimators perform worse when ρ increases; and (ii) the risks of δ_B and δ_T may be even larger than 1 when the dependence structure is strong, say for Σ_3 with $\rho \geq 0.35$.

μ	τ	n	Sample Mean	James-Stein	Berger-Bock	Tong et.al.	Proposed
μ_1	0.5	10	1.0064	0.5160	0.6175	**0.4763**	0.4867
		25	1.0020	0.8092	0.7428	0.7195	**0.7184**
		50	0.9971	0.8853	0.7865	0.7787	**0.7766**
		100	0.9954	1.0426	0.9191	0.9182	**0.9179**
	1	10	0.9996	0.9524	0.8710	**0.8266**	0.8323
		25	1.0029	0.9231	0.8937	0.8827	**0.8811**
		50	0.9963	0.9868	0.9453	0.9430	**0.9429**
		100	1.0019	1.0024	0.9802	0.9795	**0.9793**
μ_2	0.5	10	1.0036	0.5693	0.6446	0.5083	**0.4229**
		25	1.0010	0.8119	0.7434	0.7165	**0.6160**
		50	0.9945	0.9188	0.8338	0.8268	**0.7535**
		100	0.9940	0.9752	0.9062	0.9044	**0.8605**
	1	10	0.9983	0.8989	0.8578	0.8006	**0.7300**
		25	0.9959	0.9759	0.9155	0.9062	**0.8612**
		50	0.9919	0.9927	0.9489	0.9470	**0.9223**
		100	0.9969	0.9949	0.9720	0.9716	**0.9586**

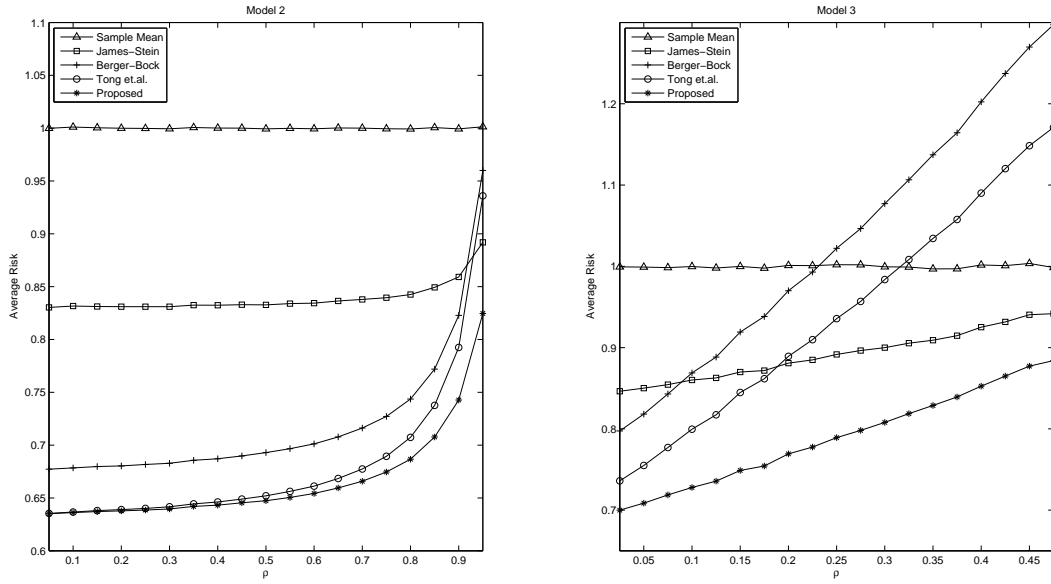Table 1: Average risks of the estimators under variance settings with covariance matrix Σ_1.

Figure 1: Plots of the average risks of the proposed method and existing methods when the observations are correlated. Here $p = 100$, $n = 20$ and $\mu = \mu_1$ with $\tau = 0.5$.

5 An application

In this section, we illustrate the proposed shrinkage estimator using the Leukemia data in Golub *et al.* (1999). The data set contains $p = 7129$ genes for 47 acute lymphoblastic leukemia (ALL) and 25 acute myeloid leukemia (AML), and is available online at the website *http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi*.

To evaluate the performance of the proposed estimator and compare it with existing methods, we randomly split the 47 ALL samples into the training set X_1 and the test set X_2. Specifically, we let the size of the training set range from 5 to 30 and the remaining samples assigned as the test set. Let $\bar{X}_1$ and $\bar{X}_2$ be the sample means of the training and test sets, respectively. We further standardize the ALL and AML sets so that each array has variance one across genes. For simplicity, we let $Q = I_p$ and let the loss function (4.10) be $L(\delta|\mu) = (n/p)(\delta - \mu)'(\delta - \mu)$. Then to compare the performance of the shrinkage estimator δ and the sample mean $\bar{X}_1$ based on the training set, we define the empirical PRIAL as

$$\text{EPR} = \frac{L(\bar{X}_2|\bar{X}_1) - L(\bar{X}_2|\delta)}{L(\bar{X}_2|\bar{X}_1)} = 1 - \frac{|\bar{X}_2 - \delta|^2}{|\bar{X}_2 - \bar{X}_1|^2}. \tag{5.11}$$

Intuitively, if δ estimates the true mean μ more accurately than the sample mean $\bar{X}_1$, it will serve as a better proctor of μ and so $L(\bar{X}_2|\delta)$ will be smaller than $L(\bar{X}_2|\bar{X}_1)$. As a consequence, if the estimated EPR is larger than 0, we may clarify that δ is better than $\bar{X}_1$. Or equivalently, the EPR may represent the improvement of δ over $\bar{X}_1$.

With $10,000$ simulations, we plot in Figure 2 the average EPR using the first 100 and 200 genes of the AML and ALL sets with different sizes of the training set. Similarly as in Section 4, it is evident that the proposed estimator $\hat{\mu}^*$ outperforms the shrinkage estimator δ_T in most settings. We also note that the improvement of $\hat{\mu}^*$ over δ_T becomes smaller when the size of the training set increases. This shows that when the sample size is large, the performance of $\hat{\mu}^*$ over δ_T will be very similar. Meanwhile, the decreasing pattern of EPR on the training size indicates that both $\hat{\mu}^*$ and δ_T reduce to the sample mean $\bar{X}_1$ when the sample size is large.
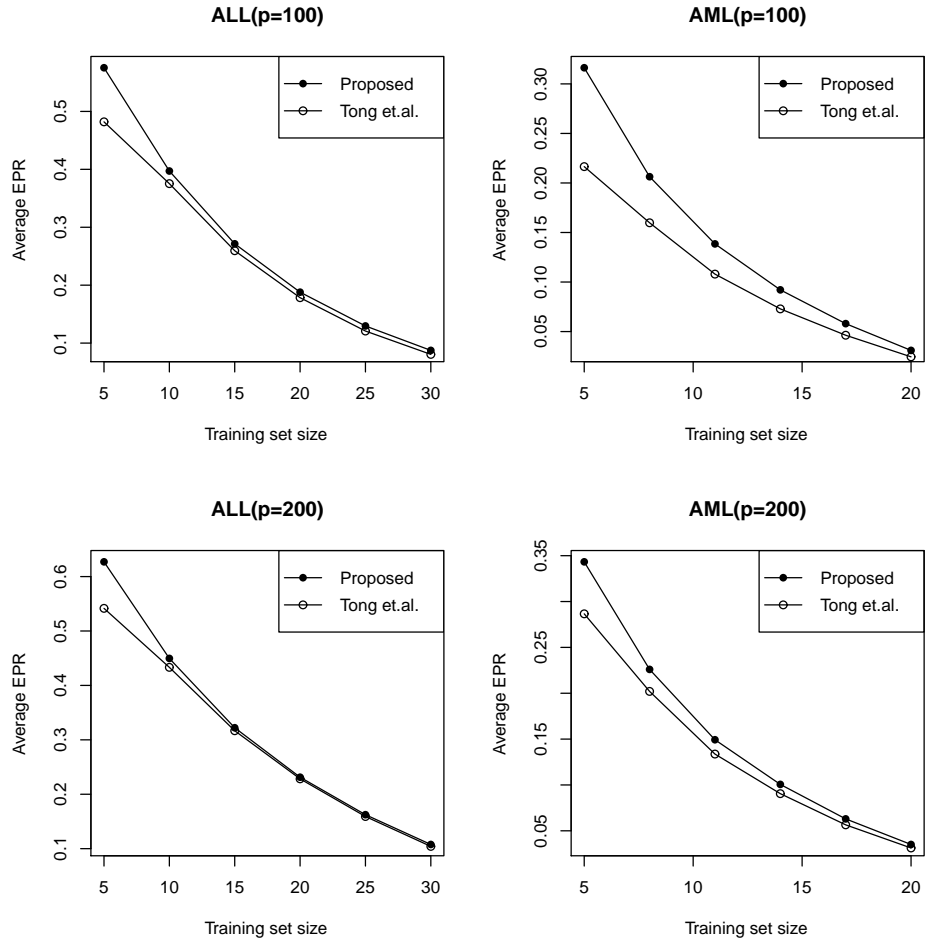
Figure 2: The average EPRs of different shrinkage estimators on Leukemia data.

6 Conclusion

The paper focuses on the shrinkage mean estimation under the "large p small n" setting. Specifically, we proposed a shrinkage estimator for the population mean under arbitrary quadratic loss functions with unknown covariance matrix. Unlike existing methods in the literature, the proposed method does not assume a specific parametric distribution for the data and does not require any prior information on the covariance matrix. In this sense, the proposed estimator is a non-parametric shrinkage estimator and it works for both diagonal and non-diagonal covariance matrices. To verify the proposed estimator, we derived some analytical results on the estimator and on the optimal shrinkage coefficients. The estimators of the optimal shrinkage coefficients were also derived along with some asymptotic properties. We have also demonstrated through simulation studies using simulated data and real data that the proposed shrinkage estimator performs better than the sample mean estimation and the existing shrinkage methods under the "large p small n" setting. Finally, we note that the proposed method (i) extends the methods in Berger *et al.* (1977) and Gleser (1986) from the "small p large n" setting to the "large p small n" setting; and (ii) extends the methods in Berger and Bock (1976) and Tong *et al.* (2012a) from a diagonal covariance matrix assumption to a non-diagonal covariance matrix assumption. The proposed method has extensive applications in different areas including statistical genetics, epidemiology, ecology, and engineering sciences.

Appendix: Proofs of the Theorems

A.1. Proof of Theorem 2.1

By direct calculation, we have

$$
\begin{aligned}
E(\delta - \mu)'Q(\delta - \mu) &= \alpha^2(\mu'Q\mu + \frac{1}{n}\mathrm{tr}(Q\Sigma_p)) + 2\alpha\mu'Q(\beta e - \mu) + (\beta e - \mu)'Q(\beta e - \mu) \\
&= \alpha^2(\mu'Q\mu + \frac{1}{n}\mathrm{tr}(Q\Sigma_p)) - (2\alpha - 1)\mu'Q\mu + \beta^2 e'Qe - 2\beta(1 - \alpha)e'Q\mu.
\end{aligned}
$$

This leads to the optimal weights as

$$\alpha^* = \frac{\mu'Q\mu - \frac{(e'Q\mu)^2}{e'Qe}}{\mu'Q\mu + \frac{1}{n}\text{tr}(Q\Sigma_p) - \frac{(e'Q\mu)^2}{e'Qe}}$$

$$\beta^* = \frac{e'Q\mu}{e'Qe}(1 - \alpha^*).$$

Further, we have

$$
\begin{aligned}
E[L_Q(\mu^*)] &= \frac{(\mu'Q\mu - \frac{(e'Q\mu)^2}{e'Qe})}{\mu'Q\mu + \frac{1}{n}\text{tr}(Q\Sigma_p) - \frac{(e'Q\mu)^2}{e'Qe}} \\
&= \frac{(\mu - \frac{e'Q\mu}{e'Qe}e)'Q(\mu - \frac{e'Q\mu}{e'Qe}e)}{(\mu - \frac{e'Q\mu}{e'Qe}e)'Q(\mu - \frac{e'Q\mu}{e'Qe}e) + \frac{1}{n}\text{tr}(Q\Sigma_p)}.
\end{aligned}
$$

A.2. Proof of Theorem 3.1

Without loss of generality, we assume that $E(\epsilon_{11}^4) = 3 + \Delta$. Then $E(Y_{1,n}) = (n/p)\mu'Q\mu$, $E(Y_{2,n}) = (1/p)\text{tr}(\Sigma_p Q)$, $E(Y_{3,n}) = n(\mu'Qe)^2/(pe'Qe)$ and $E(Y_{4,n}) = e'Qu/(e'Qe)$. In addition, we have

$$\text{Var}(Y_{1,n}) = \frac{2n}{p^2(n-1)}\text{tr}(\Sigma_p Q\Sigma_p Q) = O(\frac{1}{p}),$$

$$\text{Var}(Y_{2,n}) = \frac{2}{p^2(n-1)}\text{tr}(\Sigma_p Q\Sigma_p Q) + \frac{\Delta}{p^2 n}\text{tr}((\Sigma_p^{1/2}Q\Sigma_p^{1/2}) \circ (\Sigma_p^{1/2}Q\Sigma_p^{1/2})) = O(\frac{1}{np}),$$

$$\text{Var}(Y_{3,n}) = \frac{2n}{p^2(n-1)}(\frac{e'Q\Sigma_p Qe}{e'Qe})^2 = O(\frac{1}{p^2}),$$

$$\text{Var}(Y_{4,n}) = \frac{e'Q\Sigma_p Qe}{n(e'Qe)^2} = O(\frac{1}{np}),$$

where $A \circ B = (a_{ij}b_{ij})$ for matrices $A = (a_{ij})$ and $B = (b_{ij})$. This leads to

$$Y_{1,n} = \frac{n}{p}\mu'Q\mu + O_p(\frac{1}{\sqrt{p}}), \qquad Y_{2,n} = \frac{1}{p}\text{tr}(\Sigma_p Q) + O_p(\frac{1}{\sqrt{np}}),$$

$$Y_{3,n} = \frac{n(\mu'Qe)^2}{pe'Qe} + O_p(\frac{1}{p}), \qquad Y_{4,n} = \frac{e'Qu}{e'Qe} + O_p(\frac{1}{\sqrt{np}}).$$

16

For $\hat{\alpha}^*$ and $\hat{\beta}^*$, we have

$$
\begin{aligned}
\hat{\alpha}^* - \alpha^* &= \frac{\frac{n}{p}\pi_2 + O_p(\frac{1}{\sqrt{p}})}{\frac{n}{p}\pi_1 + \frac{n}{p}\pi_2 + O_p(\frac{1}{\sqrt{p}})} - \frac{\frac{n}{p}\pi_2}{\frac{n}{p}\pi_1 + \frac{n}{p}\pi_2} \\
&= \frac{(\frac{n}{p}\pi_2 + \frac{n}{p}\pi_1)O_p(\frac{1}{\sqrt{p}})}{(\frac{n}{p}\pi_1 + \frac{n}{p}\pi_2 + O_p(\frac{1}{\sqrt{p}}))(\frac{n}{p}\pi_1 + \frac{n}{p}\pi_2)} \\
&= O_p(\frac{1}{\sqrt{p}(\frac{n}{p}\pi_1 + \frac{n}{p}\pi_2)}) \\
&= O_p(\frac{1}{\sqrt{p}}),
\end{aligned}
$$

and

$$
\begin{aligned}
\hat{\beta}^* - \beta^* &= (1 - \hat{\alpha}^*)Y_{4,n} - \beta^* \\
&= (1 - \alpha^*)(Y_{4,n} - \frac{e'Qu}{e'Qe}) - (\hat{\alpha}^* - \alpha^*)Y_{4,n} \\
&= O_p(\frac{1}{\sqrt{np}})) + (\frac{e'Qu}{e'Qe} + O_p(\frac{1}{\sqrt{np}}))O_p(\frac{1}{\sqrt{p}(\frac{n}{p}\pi_1 + \frac{n}{p}\pi_2)}) \\
&= O_p(\frac{1}{\sqrt{np}}) + O_p(\frac{e'Qu}{p^{3/2}(\frac{n}{p}\pi_1 + \frac{n}{p}\pi_2)}).
\end{aligned}
$$

A.3. Proof of Theorem 3.2

First consider $(\mu^* - \mu)'Q(\mu^* - \mu)$. Note that

$$
\begin{aligned}
&(\mu^* - \mu)'Q(\mu^* - \mu) \\
=\ & (\alpha^*(\bar{X} - \mu) + \beta^*e + (\alpha^* - 1)\mu)'Q(\alpha^*(\bar{X} - \mu) + \beta^*e + (\alpha^* - 1)\mu) \\
=\ & (\alpha^*)^2(\frac{1}{n^2}\sum_{k=1}^{n}\epsilon_k'\Sigma_p^{1/2}Q\Sigma_p^{1/2}\epsilon_k + \frac{1}{n^2}\sum_{i\neq j}\epsilon_i'\Sigma_p^{1/2}Q\Sigma_p^{1/2}\epsilon_j) \\
& + \frac{2\alpha^*}{n}(\beta^*e + (\alpha^* - 1)\mu)'Q\Sigma_p^{1/2}\sum_{k=1}^{n}\epsilon_k + (\beta^*e + (\alpha^* - 1)\mu)'Q(\beta^*e + (\alpha^* - 1)\mu).
\end{aligned}
$$

17

Then,

$$Var((\mu^* - \mu)'Q(\mu^* - \mu)) \leq 2(\alpha^*)^4 [\frac{2}{n^2} tr(\Sigma_p Q \Sigma_p Q) + \frac{\Delta}{n^3} \sum_{k=1}^{p} \phi_{kk}^2]$$
$$+ \frac{8(\alpha^*)^2(1-\alpha^*)^2}{n}(\mu - \frac{e'Q\mu}{e'Qe}e)'Q\Sigma_p Q(\mu - \frac{e'Q\mu}{e'Qe}e),$$

where $\Sigma_p^{1/2} Q \Sigma_p^{1/2} = (\phi_{ij})_{p \times p}$.

By the definitions of π_1 and π_2, it is easy to verify that

$$\alpha^* = \frac{\pi_2}{\pi_1 + \pi_2},$$

$$E(\mu^* - \mu)'Q(\mu^* - \mu) = \frac{\pi_1 \pi_2}{\pi_1 + \pi_2},$$

$$Var((\mu^* - \mu)'Q(\mu^* - \mu)) \leq \frac{C_1}{n} \frac{\pi_1 \pi_2^4 + \pi_1^2 \pi_2^3}{(\pi_1 + \pi_2)^4} = (\frac{\pi_1 \pi_2}{\pi_1 + \pi_2})^2 O(\frac{1}{p}).$$

Therefore,

$$(\mu^* - \mu)'Q(\mu^* - \mu) = \frac{\pi_1 \pi_2}{\pi_1 + \pi_2}(1 + O_p(\frac{1}{\sqrt{p}})). \qquad (6.12)$$

By Theorem 3.1, we have

$$\hat{\mu}^* = \mu^* + O_p(\frac{1}{\sqrt{p}})\bar{X} + (O_p(\frac{1}{\sqrt{np}}) + O_p(\frac{e'Qu}{p^{3/2}}))e.$$

Note that

$$\begin{aligned}
\hat{\beta}^* - \beta^* &= (1 - \hat{\alpha}^*)Y_4 - (1 - \alpha^*)\frac{e'Qu}{e'Qe} \\
&= (\alpha^* - \hat{\alpha}^*)\frac{e'Qu}{e'Qe} + (1 - \hat{\alpha}^*)(Y_4 - \frac{e'Qu}{e'Qe}), \\
\hat{\beta}^* + \beta^* &= (1 - \hat{\alpha}^*)Y_4 + (1 - \alpha^*)\frac{e'Qu}{e'Qe} \\
&= (2 - \alpha^* - \hat{\alpha}^*)\frac{e'Qu}{e'Qe} + (1 - \hat{\alpha}^*)(Y_4 - \frac{e'Qu}{e'Qe}).
\end{aligned}$$

18

We have

$$(\hat{\mu}^* - \mu)'Q(\hat{\mu}^* - \mu) - (\mu^* - \mu)'Q(\mu^* - \mu)$$

$$= (\hat{\mu}^* - \mu^*)'Q(\hat{\mu}^* + \mu^* - 2u)$$

$$= ((\hat{\alpha}^* - \alpha^*)\bar{X} + (\hat{\beta}^* - \beta^*)e)'Q((\hat{\alpha}^* + \alpha^*)\bar{X} + (\hat{\beta}^* + \beta^*)e - 2u)$$

$$= ((\hat{\alpha}^* - \alpha^*)(\bar{X} - \mu) + (\alpha^* - \hat{\alpha}^*)(\frac{e'Qu}{e'Qe}e - \mu) + (1 - \hat{\alpha}^*)(Y_4 - \frac{e'Qu}{e'Qe})e)'Q$$

$$\quad *((\hat{\alpha}^* + \alpha^*)(\bar{X} - \mu) + (2 - \hat{\alpha}^* - \alpha^*)(\frac{e'Q\mu}{e'Qe}e - \mu) + (1 - \hat{\alpha}^*)(Y_4 - \frac{e'Qu}{e'Qe})e)$$

$$= (\hat{\alpha}^* - \alpha^*)[(\hat{\alpha}^* + \alpha^*)(\bar{X} - \mu)'Q(\bar{X} - \mu) + 2(1 - \hat{\alpha}^* - \alpha^*)(\bar{X} - \mu)'Q(\frac{e'Q\mu}{e'Qe}e - \mu)]$$

$$\quad + 2\hat{\alpha}^*(1 - \hat{\alpha}^*)(Y_4 - \frac{e'Qu}{e'Qe})(\bar{X} - \mu)'Qe + (Y_4 - \frac{e'Qu}{e'Qe})^2(1 - \hat{\alpha}^*)^2 e'Qe$$

$$\quad + (\hat{\alpha}^* - \alpha^*)(2 - \hat{\alpha}^* - \alpha^*)(\mu - \frac{e'Qu}{e'Qe}e)'Q(\mu - \frac{e'Qu}{e'Qe}e)$$

$$= \pi_1 O_p(\frac{1}{\sqrt{p}}) + O_p(\frac{1}{\sqrt{p}(\frac{n}{p}\pi_1 + \frac{n}{p}\pi_2)})O_p(\sqrt{\frac{\pi_2}{n}}) + O_p(\frac{1}{n}) + O_p(\frac{\pi_2}{\sqrt{p}(\frac{n}{p}\pi_1 + \frac{n}{p}\pi_2)})$$

$$= \pi_1(O_p(\frac{1}{\sqrt{p}}) + O_p(\frac{1}{p})) = \pi_1 O_p(\frac{1}{\sqrt{p}}), \qquad (6.13)$$

where we used the facts that

$$(\bar{X} - \mu)'Q(\bar{X} - \mu) = \pi_1(1 + O_p(\frac{1}{\sqrt{p}})),$$

$$(\bar{X} - \mu)'Q(\frac{e'Q\mu}{e'Qe}e - \mu) = O_p(\sqrt{\frac{\pi_2}{n}}),$$

$$(\bar{X} - \mu)'Qe = O_p(\sqrt{\frac{p}{n}}),$$

and

$$\hat{\alpha}^* - \alpha^* = O_p(\frac{1}{\sqrt{p}(\frac{n}{p}\pi_1 + \frac{n}{p}\pi_2)}).$$

Finally, by (6.12) and (6.13) we have

$$(\hat{\mu}^* - \mu)'Q(\hat{\mu}^* - \mu) = \pi_1(\frac{\pi_2}{\pi_1 + \pi_2} + O_p(\frac{1}{\sqrt{p}})). \qquad (6.14)$$

This completes the proof of Theorem 3.2.

19

Acknowledgments

References

Anderson, T. (2003). *An Introduction to Multivariate Statistical Analysis.* Wiley Series in Probability and Statistics.

Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, **6**, 311–330.

Baldi, P. and Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509–519.

Baranchik, A. (1970). A family of minimax estimators of the mean of a multivariate normal distribution. *The Annals of Mathematical Statistics*, **41**, 642–645.

Berger, J. (1976). Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss. *The Annals of Statistics*, **4**, 223–226.

Berger, J. and Bock, M. (1976). Combining independent normal mean estimation problems with unknown variances. *The Annals of Statistics*, **4**, 642–648.

Berger, J., Bock, M., Brown, L., Casella, G., and Gleser, L. (1977). Minimax estimation of a normal mean vector for arbitrary quadratic loss and unknown covariance matrix. *The Annals of Statistics*, **5**, 763–771.

Cai, T. and Ma, Z. (2012). Optimal hypothesis testing for high dimensional covariance matrices. *arXiv preprint arXiv:1205.4219*.

Cai, T. and Yuan, M. (2012). Adaptive covariance matrix estimation through block thresholding. *The Annals of Statistics, in press*.

Chen, S., Zhang, L., and Zhong, P. (2010). Tests for high-dimensional covariance matrices. *Journal of the American Statistical Association*, **105**, 810–819.

Cui, X., Hwang, J. T. G., Qiu, J., Blades, N. J., and Churchill, G. A. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, **6**, 59–75.

Dudoit, S., Fridlyand, J., and Speed, T. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, **97**, 77–87.

Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitorsan empirical bayes approach. *Journal of the American Statistical Association*, **68**, 117–130.

El Karoui, N. (2010). High-dimensionality effects in the markowitz problem and other quadratic programs with linear constraints: risk underestimation. *The Annals of Statistics*, **38**, 3487–3566.

Fan, J. and Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *The Annals of Statistics*, **36**, 2605–2637.

Fan, J., Feng, Y., and Tong, X. (2012). A ROAD to classification in high dimensional space. *Journal of the Royal Statistical Society, Series B*, **74**, 745–771.

Fourdrinier, D., Strawderman, W., and Wells, M. (2003). Robust shrinkage estimation for elliptically symmetric distributions with unknown covariance matrix. *Journal of Multivariate Analysis*, **85**, 24–39.

Gleser, L. (1986). Minimax estimators of a normal mean vector for arbitrary quadratic loss and unknown covariance matrix. *The Annals of Statistics*, **14**, 1625–1633.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., *et al.* (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Hwang, J., Qiu, J., and Zhao, Z. (2009). Empirical bayes confidence intervals shrinking both means and variances. *Journal of the Royal Statistical Society, Series B*, **71**, 265–285.

Hwang, J. T. G. and Liu, P. (2010). Optimal tests shrinking both means and variances applicable to microarray data analysis. *Statistical Applications in Genetics and Molecular Biology*, **9**, 36.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, **1**, 361–379.

Ledoit, O. and Wolf, M. (2004a). Honey, I shrunk the sample covariance matrix. *Journal of Portfolio Management*, **30**, 110–119.

Ledoit, O. and Wolf, M. (2004b). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, **88**, 365–411.

Li, J. and Chen, S. (2012). Two sample tests for high-dimensional covariance matrices. *The Annals of Statistics*, **40**, 908–940.

Lin, P. and Tsai, H. (1973). Generalized bayes minimax estimators of the multivariate normal mean with unknown covariance matrix. *The Annals of Statistics*, **1**, 142–145.

Lindley, D. (1962). Discussion of professor steins paper: confidence sets for the mean of a multivariate normal distribution. *Journal of the Royal Statistical Society, Series B*, **24**, 285–287.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, **7**, 77–91.

Pang, H., Tong, T., and Zhao, H. (2009). Shrinkage-based diagonal discriminant analysis and its applications in high-dimensional data. *Biometrics*, **65**, 1021–1029.

Pourahmadi, M. (2011). Covariance estimation: the GLM and regularization perspectives. *Statistical Science*, **26**, 369–387.

Ramey, J. (2012). Selected topics in high-dimensional statistical learning. *http://hdl.handle.net/2104/8513*.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale co-variance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, **4**, 32.

Shao, J., Wang, Y., Deng, X., and Wang, S. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of Statistics*, **39**, 1241–1265.

Shinozaki, N. (1980). Estimation of a multivariate normal mean with a class of quadratic loss functions. *Journal of the American Statistical Association*, **75**, 973–976.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiment. *Statistical Applications in Genetics and Molecular Biology*, **3**, 1.

Tong, T. and Wang, Y. (2007). Optimal shrinkage estimation of variances with applications to microarray data analysis. *Journal of the American Statistical Association*, **102**, 113–122.

Tong, T., Chen, L., and Zhao, H. (2012a). Improved mean estimation and its application to diagonal discriminant analysis. *Bioinformatics*, **28**, 531–537.

Tong, T., Jang, H., and Wang, Y. (2012b). James-Stein type estimators of variances. *Journal of Multivariate Analysis*, **107**, 232–243.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, **98**, 5116–5121.

Van der Vaart, A. (2000). *Asymptotic Statistics*. Cambridge University Press.